**Proposal**

**Introduction**

In this project, we propose to investigate the problem that how to identify the different copyists in the “Avila Bible”by using Decision tree model. The dataset is from UC Irvine machine learning open dataset. The Avila data set has been extracted from 800 images of the the "Avila Bible", a giant Latin copy of the whole Bible produced during the XII century between Italy and Spain. The palaeographic analysis of the manuscript has individuated the presence of 12 copyists. The pages written by each copyist are not equally numerous.

**Data Processing**

The dataset has 20867 instances. a training set containing 10430 samples, and a test set containing the 10437 samples. In each data sample, the sample has 10 variables. They are intercolumnar distance, upper margin, lower margin, exploitation, row number, modular ratio, interlinear spacing, weight, peak number, modular ratio/interlinear spacing. For the response variable, there are 12 class, which means there are 12 copyists.

We plan to do some data visualization on the different variables like boxplots, correlation plots and other related plot to mine the inner trends of the data.

**Model**

For the approach part, we choose to use decision tree model to conduct our classification problem. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. One of an advantage of the decision tree model is that after training the data , we can have an intuitive analysis on all attributes. We can see which attribute is more important, which is not, by looking at the node position of the decision tree. If one attribute is close to the root node, which means that the attribute is important and we should consider this variable earlier. Otherwise, if one attribute is close to leaf, which means that the attribute is less important.

We plan to split the training data into 10 folds to conduct the cross validation. For the final results, we will report the F score of the test data. Moreover, we will compare different hyperparameters of the decision tree to see the influence of the F score. If time permits, we will also compare different machine learning classification methods such logistic regression, to see difference between the models.